Advanced
Livermore Computing Resource Management System Training

Don Lipari

Lawrence Livermore National Laboratory

lipari@llnl.gov

UCRL-WEB-202810

Introduction

- A more detailed look into the workings of LCRM
- Provides answers to common questions
- Will help to diagnose job scheduling difficulties

Agenda

- Nomenclature
- LCRM vs. SLURM / LoadLeveler
- How LCRM Works
- Job Information
- Commands
- LCRM Libraries
- Common Concerns
- Information Resources

Nomenclature

- Accounts, banks, and user permissions
- Multi-node vs. SMP machines
- Constraints and Limits
- Features
- Interactive vs. LCRM jobs
- LCRM jobs: normal, expedited, standby, and delayed

Nomenclature (cont.)

- Node geometry
- Partitions and pools, LCRM pools and resource partitions
- Priority: scheduling vs. running
- Resources: CPU time and memory
- Shares, usage, and quotas
- Wall clock vs. task time

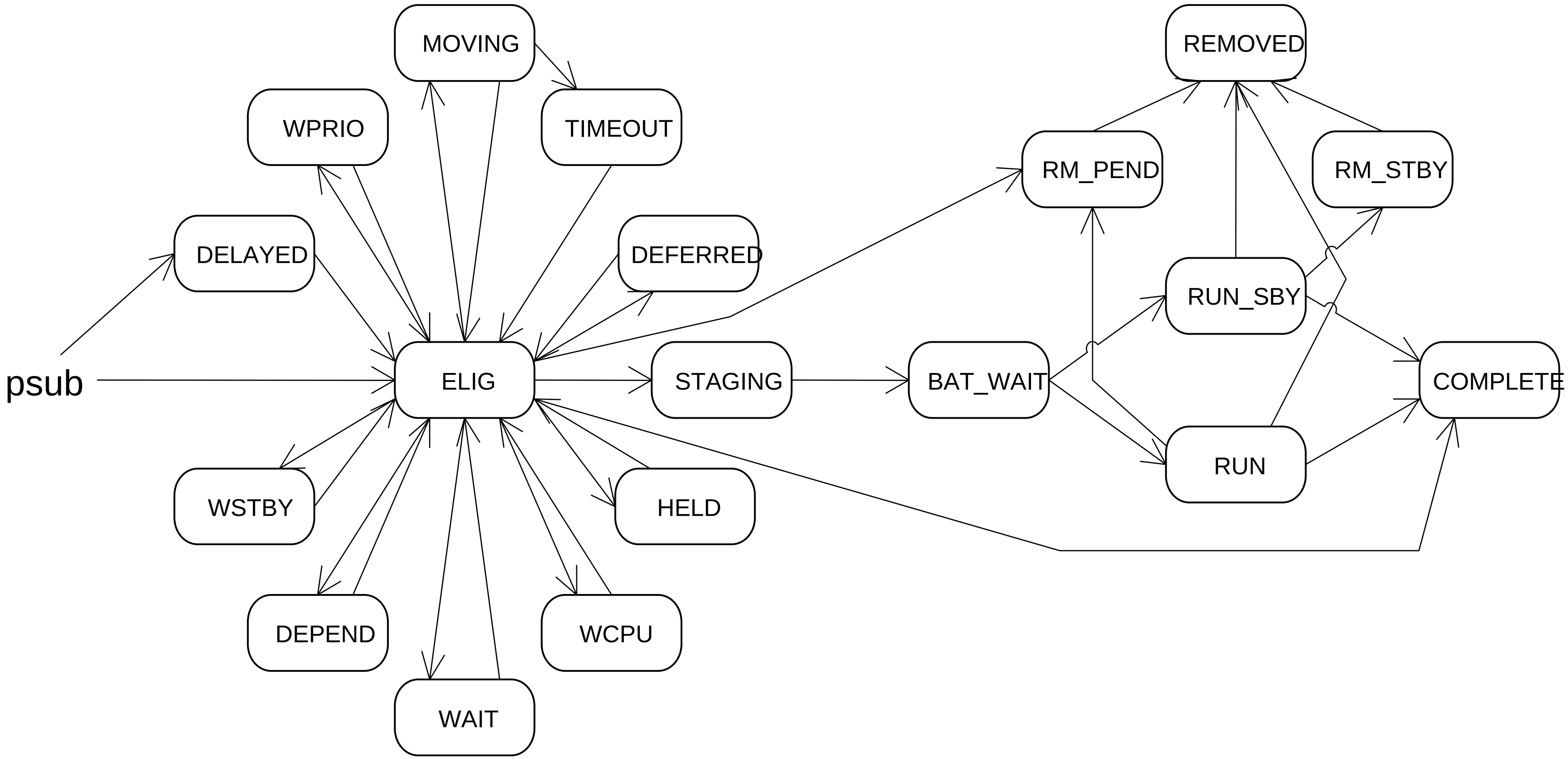
LCRM vs. SLURM / LoadLeveler

- SLURM and LoadLeveler are the native batch systems for multi-node hosts
- LCRM depends on SLURM/LL to launch LCRM jobs on multi-node hosts
- Interactive jobs cannot usually run on the largest node partitions/pools
- Debug runs are commonly launched interactively
- Interactive resource usage is charged to the default bank

How LCRM Works

- Daemons
- Scheduling (node, cluster, and memory)
- Priority (fair-share, aging, technical)
- Backfill Scheduling
- Accounting
- Signaling, nicing, and terminating

LCRM Job State Transitions



LCRM User Commands

- Job related: psub, pstat, spj, spjstat, palter, prm, pexp, phold, prel, phist
- Account related: acc, bac, defacct, defbank, newacct, newbank, pshare
- Limit displays: brlim, plim, pquota

Not LCRM User Commands

- sinfo, squeue, llq, jj, ju, etc.

psub

- Can be invoked on any machine within the LCRM domain to run on any host
- A copy of the job script is spooled to the submitting host at submit time
- Usually, invalid options are rejected at submit time
- Supported shells: bourne, csh, perl
- Accommodated shells: bash, korn, tcsh
- news tcsh.batch

Geometry Related psub Options

Scheduler Type	**-cpn**	**-g**	**-ln**
Node - LoadLeveler	Ignored	Optional	Optional
Node - SLURM	Optional	Ignored	Required
Node – RMS	Optional	Ignored	Required*
Cluster	Recommended Used by Scheduler	Ignored	Ignored
Memory	Ignored	Ignored	Ignored

* -ln heterogeneous job support on RMS is not available

pstat

- You can define your favorite fields to be listed in the PSTAT_CONFIG env variable
- pstat –f gives you the most detailed report
- pstat –M reveals the multiple reasons for the MULTIPLE job state

LCRM Administrator Commands

- lrmmgr
- lrmmon
- palter
- pcsusage
- phstat
- pundelay

Email From LCRM

- Resource limit has been reached
- Limit change affects job
- Administrator removed a required host
- Standby job was removed for priority job
- Miscellaneous system errors
- Error messages directly from batch system

LCRM Libraries

- lrmsig_register
- lrmgetresource / lrmgettime
- lrmwarn
- When called multiple times, last call takes effect
- Library: /usr/local/lib/liblrm.a
- Include: /usr/local/include/liblrm.h

Common Concerns

- Why isn’t my job running?
- Why did my job terminate?
- What banks do I belong to?
- What banks should I use?

Reasons for Not Starting Job

pstat STATUS	**LCRM Host Config**	**plim**	**User/bank partition**	**brlim**	**psub Arguments / current conditions**
CPU&TIME	maxnodetime	-nh Max. allowable node-hours for running batch jobs			-ln * -tW
CPUS>MAX	maxnodecount	-ln Max. allowable nodes for running batch jobs			-ln
DEFERRED					Invalid –ln spec
JRESLIM			maxjobsinpart	JOBS LIMIT	Number of running jobs
NRESLIM			maxnodes	NODES LIMIT	-ln
NTRESLIM			maxresrcinpart	NODE-TIME LIMIT	-ln * -tW
PTOOBIG	maxprocsize	-ms Max. allowable size for running batch jobs			-lM

More Reasons for Not Starting Job

pstat STATUS	Host Config	plim	User/bank partition	brlim	psub Arguments / current conditions
QTOTLIM	maxacttot – max active jobs total				Number of running jobs on host
QTOTLIMU	maxactuser - max active jobs per user				Number of user's running jobs on host
TOOLONG	maxcputime defcputimelim maxwalltime Defwalltimelim	-mr Maximum cpu time for batch jobs -tM Default cpu time limit for a batch job -mR Maximum run time for batch jobs -tW Default elapsed time limit for a batch job	maxjobtime		-tM –tW
WMEML(oad)	maxmemthresh maxswapthresh				-lM
WMEMT(arget)	minmemthresh minswapthresh				-lM

Accurate Time Estimates (-tM and –tW) are Important

- For low priority jobs because it determines backfill eligibility
- For high priority jobs so that next high priority job runs as soon as possible
- For high priority jobs because an inflated committed time will lower priority
- Memory and cluster scheduled machines are subject to the same limitations

Information Resources

- LCRM man pages
- LC LCRM Tutorial

  http://www.llnl.gov/computing/tutorials/lcrm/
- LCRM Reference manual

  http://www.llnl.gov/LCdocs/dpcs/
- /usr/local/docs (sample scripts)
- Technical Bulletins (304, 320, 336, …)

  https://lc.llnl.gov/computing/techbulletins/bulletin336.html